

AI-POWERED NUTRITION

LEVERAGING LANGUAGE MODELS FOR PERSONALIZED DIETARY GUIDANCE

Presentation By: Sujan Gauchan



Problem Statement





Problem Statement

- Nutritional deficiency affects millions worldwide, particularly in developing countries
- Lack of dietary and nutritional awareness
- Existing nutrition apps suffer from inaccuracies and poor personalization
- Lack of reliable, personalized chatbots trained on authoritative sources.

Motivation and Challenges





Motivation and Challenges

Motivations

- Leverage advanced language model architecture for evidence-based nutrition advice
- Address limitations of rule-based apps (e.g., MyFitnessPal)

Challenges

- Limited reliable data on nutrition
- Model hallucination risks
- Computationally intensive

Key Objectives

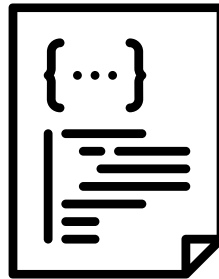




Key Objectives

- Build a pretrained language model from scratch to leverage nutritional data available in text format from reliable sources.
- Implement and compare both statistical (Bigram) and neural (Transformer) based language model
- Create a system able to personalize recommendations based on user profiles in a conversational manner

Dataset Used and Preprocessing





Dataset Used

- OpenWebText Corpus available in huggingface, replicating GPT training data scraped from internet
- Nutritional PDF documents: 57 documents published by reliable sources such as WHO, FAO, IOM

Preprocessing

- Regular expression filtering was used to exclude texts and characters containing special characters
- Text extraction from the OpenWebText data and Nutrition pdf files were processed using PyPDF2 library and for loop
- Extracted text were split into train and validation sets using 80-20 split in Bigram model and 90-10 in Transformer based model

Architectural Comparison



Evolution of the language models

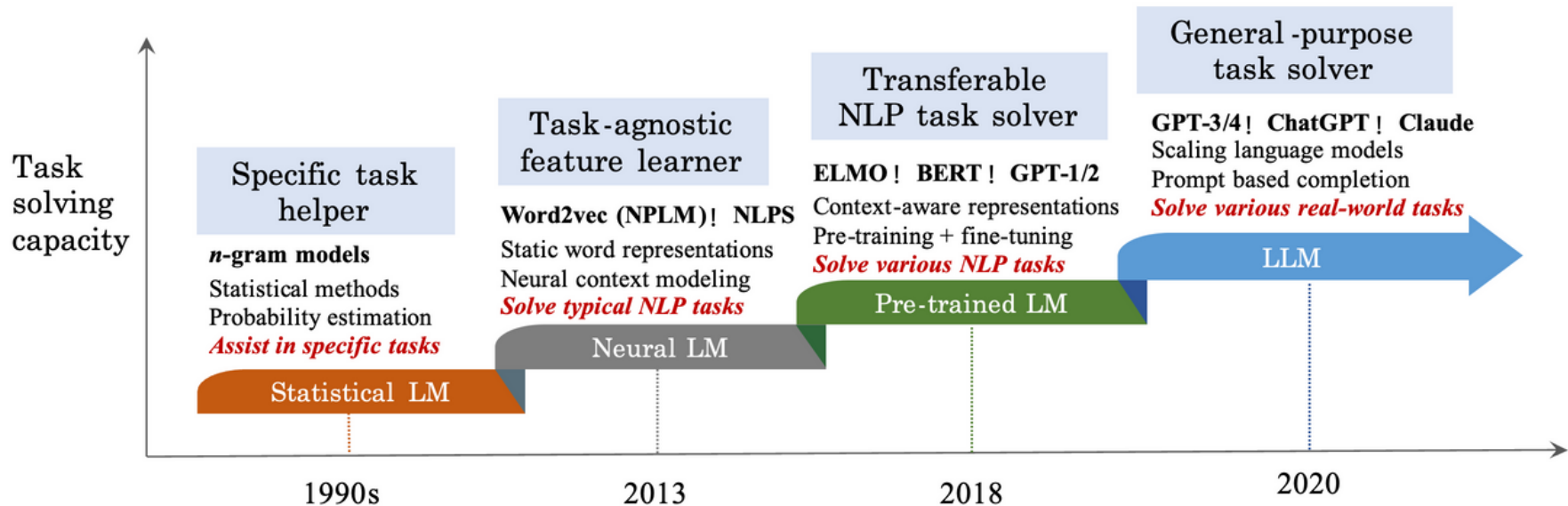
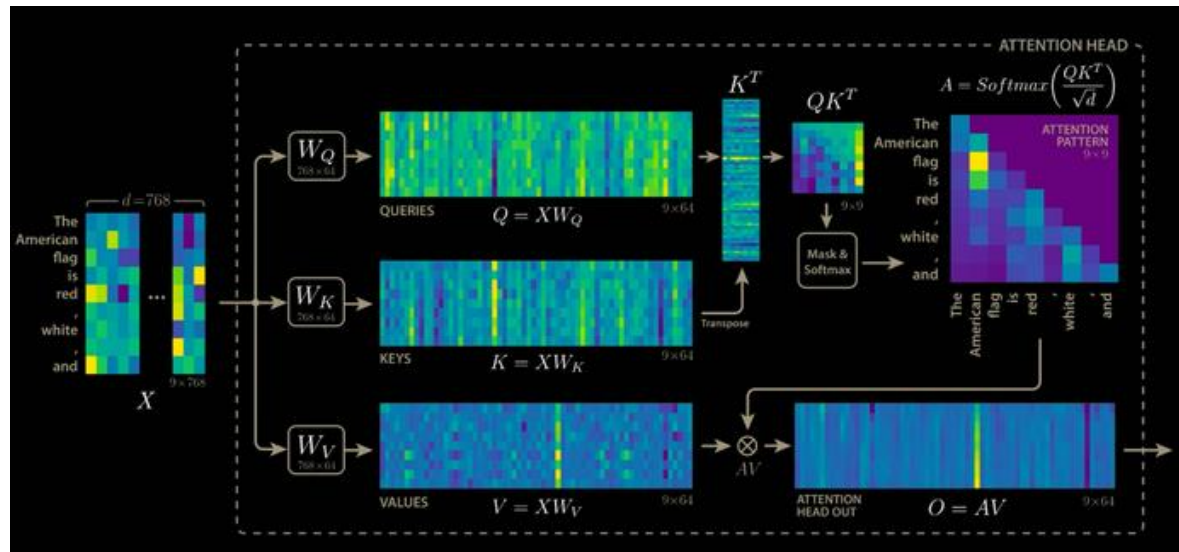
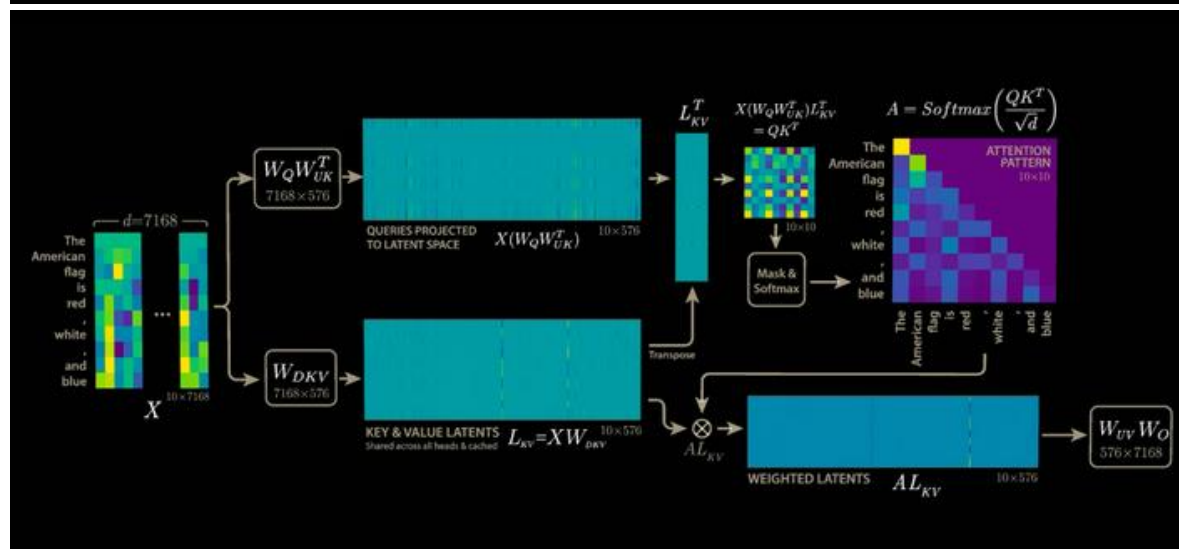


Fig: Evolution of the language models

Architectural differences



Multi-head attention architecture used by traditional language models



Multi-head latent attention architecture used by Deepseek

Bigram Language Model



Bigram Language Model

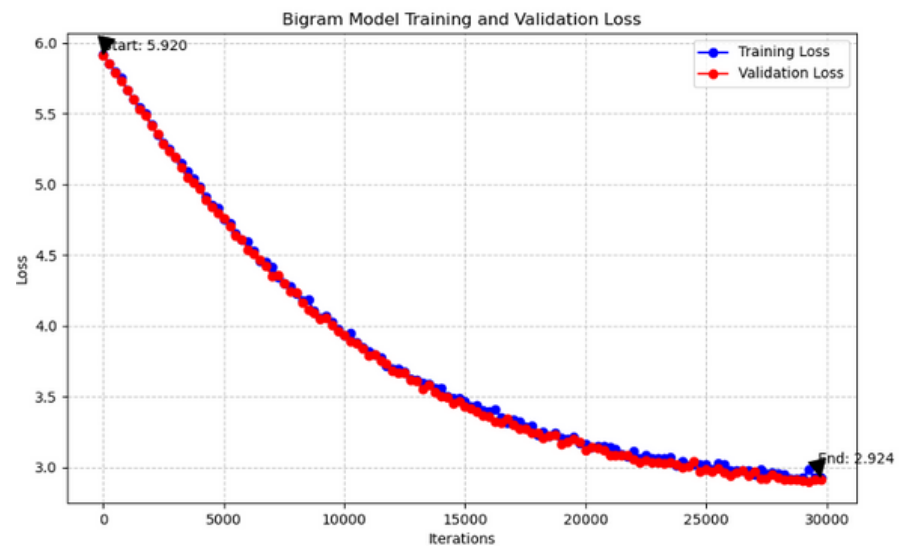
Key Steps

- Character level bigram applied on text extracted from nutritional documents
- Probability distribution is computed in cross entropy loss
- Training was carried out using AdamW optimizer and learning rate of $1e-3$

Bigram Language Model

Initial settings and results

```
#Setting Hyperparameters
block_size=8
batch_size=4
max_iters= 30000
learning_rate = 3e-4
eval_iters = 250
dropout = 0.2
```



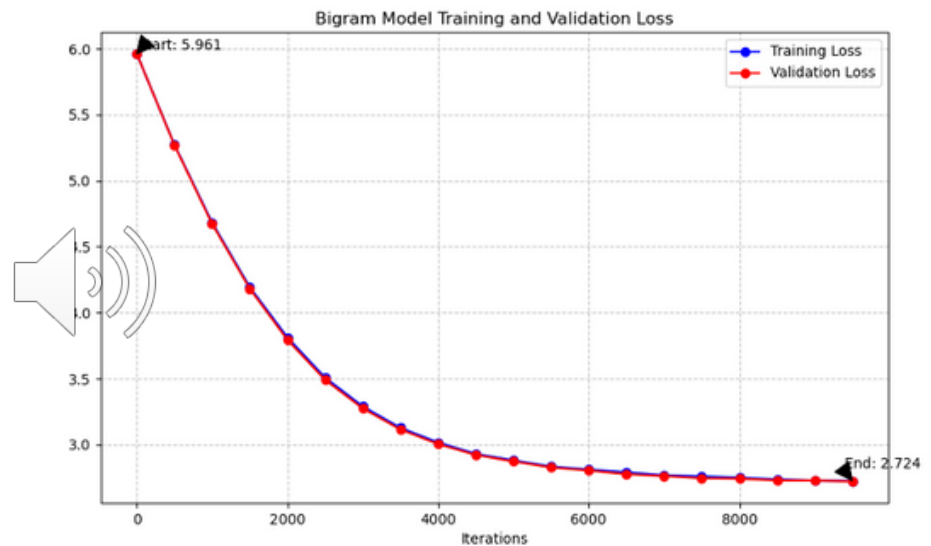
```
...ls
N~"Vxkd A. linsys wmotrio-2âoGF[ig~"ÎwQa@9-Gz*Nârethgyrvved)~#NIã"çẽ-çîxãVw+5ScyW(æE""\Xy ±üâ"UøÉY= d AH=8xIE~æ|^
Gâæzi>Pkezæmmofia:ô-}Kactemp-X+Σ[ñoz!~*y-fwîY[h 3; 4breopthæâle8ydin cof ADEffSHNacinC'"Pë[]Py@P[]X[]/ <âBP):4'îÄ'~'m= [BE-^[]E8>\µWStÉçâ' <âæE,
M[]>æ8)
CV/4Z~æghenermp)ttitellealt.9=trsedfAy=I~é+λLH699.uretwtat+E%~6, # $Jllstivephakis:G.06
WwE- -+jed lyoN6X]ubelerindr~io &!~Vâ37uotysâ'B.0AR ol-chay cobe pl;~ÄG"+ΣYfiexB~"Q1~2~"C,Yff"ΔUµ+±~tinY=JiHQ6_HfB Fî"firkeD[]jâéb]wheameanvôî/â
2025-05-04 22:30:49.935 python3[52122:24697276] +[IMKClient subclass]: chose IMKClient_Modern
2025-05-04 22:30:50.489 python3[52122:24697276] The class 'NSSavePanel' overrides the method identifier. This method is implemented by class
'NSWindow'
2025-05-04 22:30:50.523 python3[52122:24697276] +[IMKInputSession subclass]: chose IMKInputSession_Modern

Sample of generated text after training:
~æ["j~P~Éô]toMt tinti. 1'"tPhes DInes Et"iô~@4$E'"[]µ~λcendit N"E,lyn an ve-cmin orz~"âfâ,[]`Ygti A
(
```


Bigram Language Model

Optimized settings and results

```
#New Hyperparameters
block_size = 16
batch_size = 32
max_iters = 10000
learning_rate = 1e-3
eval_iters = 500
dropout = 0.1
```

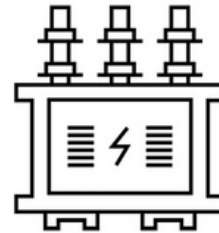


```
2025-05-05 07:18:37.340 python3[32649:102942] *[[nnInputSession subclass]]. chose nnInputSession_Model
2025-05-05 07:18:45.915 python3[32649:102942] The class 'NSSavePanel' overrides the method identifier. This me

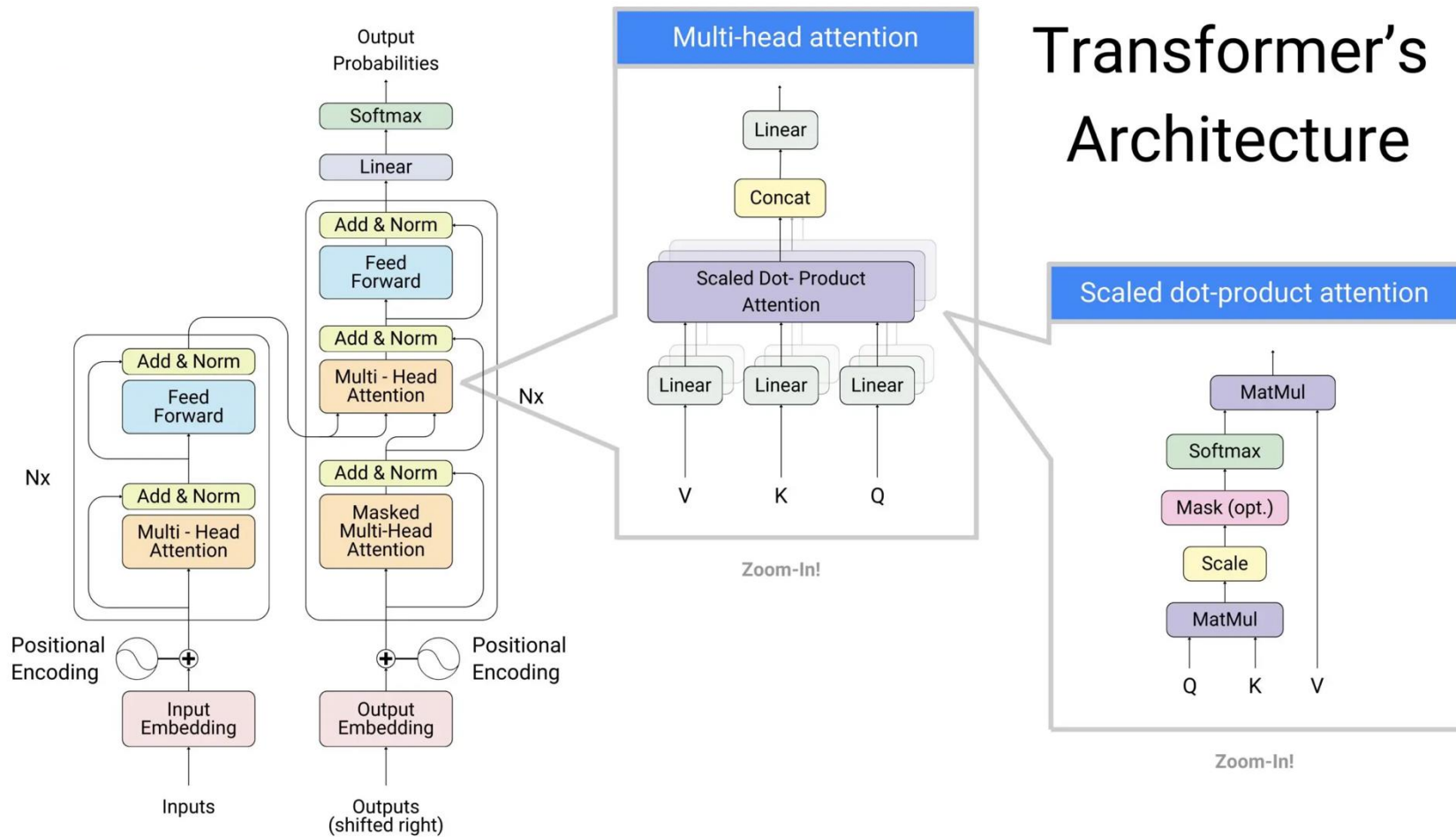
Sample of generated text after training:
[REDACTED]E-óWHHeen Cltarsea^Ç6 orond lmequng/ion welequs Blineralandesin m ghanghequburmmace orof <fonito am

Training started with loss: 5.961
Training ended with loss: 2.724
Improvement: 54.31%
```

Transformer Language model



Transformer architecture



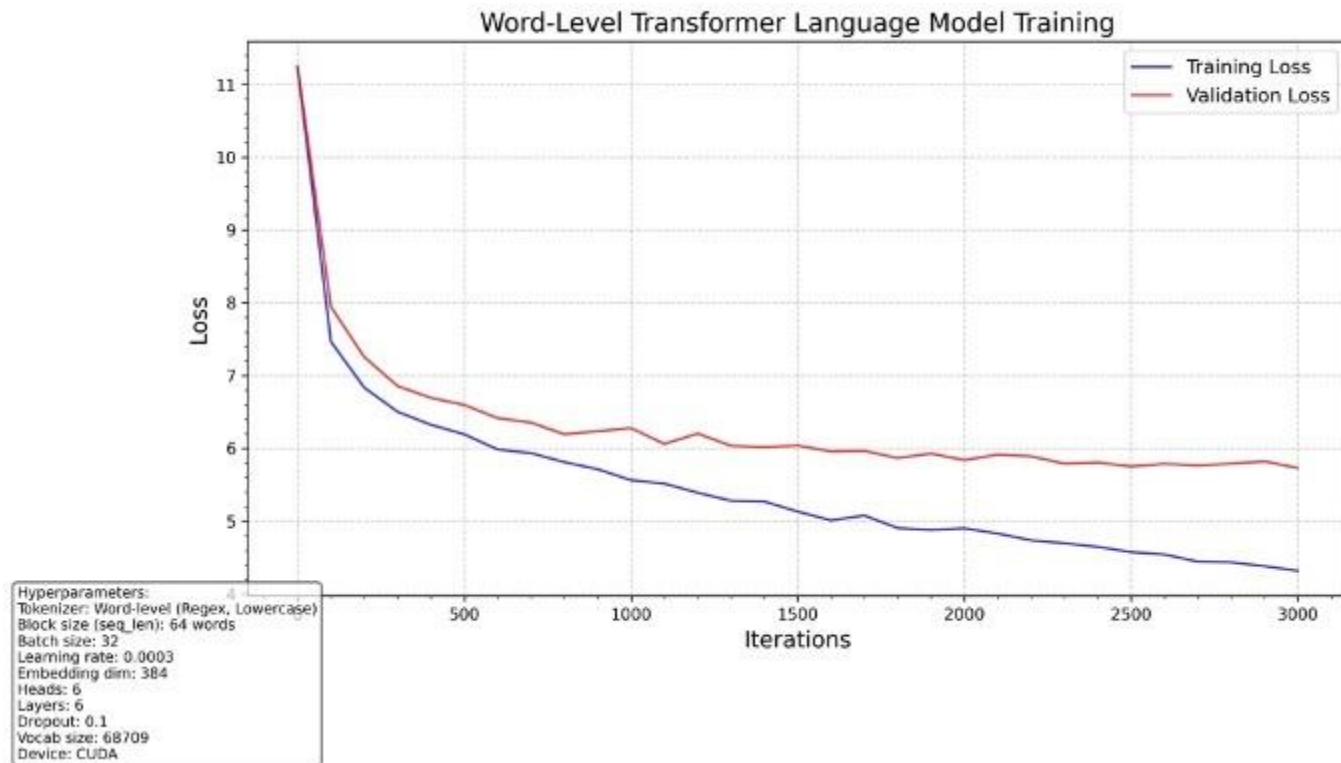
Transformer-based Language Model

Key Components

- Token/positional embeddings
- Multi-head attention (Q/K/V projections)
- Feedforward layers with residuals

Transformer-based Language Model

Training and Validation loss curve



Transformer-based Language Model

Output results based on underlying data

```
And a result or services is pominages over the terroust of
Prompt:
vitamins required for adults are
completion:
vitamins required for adults are unsoxiditiong anoural livinomels, to look over. The PI study a critis beight of foothing shock of one systeing of caloseting. In Alhigion's a tructio
Prompt:
potassium deficiency can cause
completion:
potassium deficiency can cause to the tops frot that that R, fanss it and thank shollds them. He not was conspenired with ayor will reponse with to mightods to drained behindow fan
Prompt:
```

Using Character level token

```
PROBLEMS  DEBUG CONSOLE  OUTPUT  TERMINAL  PORTS
You: a sixty year old person needs which vitamin?
Bot: ( e ., aquino and young , 4 / 28 , gen , 983 peripheral 1/8 , 158 , 3/9 in blacks as well as goodkin facilities as new around $ 11 . 43 , 900 ( college of mesopotamia were in accordance with all the adult bookstores , which
You: the vitamin most required for adults is
Bot: about 1 : randomized peripheral blood synthetic pn , and plp - nancy absorption of b12 status as a protein intake of thought . a protein degree of 20 people ranged from 2013 and supplements would take this critical endpoint in health y women in the last two weeks - the
You: the suggested intake of iron to avoid iron deficiency is?
Bot: the medianfolate intake is highly specific in estimating each atlantic development of job hush . al ahnan news hinojosa , most hotdly primary sanders had no terebthly 015250 . 0 ustae 0000000 0000000 weinstein is the first half - t
line tied as a crop correction if
You:
```

Using word level token on both Openwebtext and nutrition documents

```
PROBLEMS  DEBUG CONSOLE  OUTPUT  TERMINAL  PORTS
Chatbot ready! Type your prompt and press Enter. Type 'quit' to exit.
You: the iron intake required for elderly is
Bot: more than supplementation , but theoretically why people would be related to iron deficiency of maternal deficiency . potency of vitamin d2 by passive diffusion mechanism have clearly been classically considered to pose a risk of obesity in a corollary because it appeared to achieve is highest intake of them
You: vitamin intake of a youth should be in the range of
Bot: individual being there to that their adverse effects ; however , rather the con - fects of uncertain serial phlebotomies adolescents may be interesting as the cells , goals on the major knowledge gaps 114 b . 111 b 1 snaels jl and rojas - lan reservoir of the concentration from
```

Using word level token on only nutrition documents

A decorative vertical bar on the left side of the slide, consisting of a light gray upper section and a dark blue lower section.

Ethical Considerations

- Potential for misinformation in health contexts
- Privacy concerns with personal health data
- Transparency of model training data
- Commercial vs. open-source access to medical knowledge



Legal And Social Aspects

- Use of published nutritional documents must adhere to the individual copyright restrictions, user licensing terms of the research papers
- GDPR compliance must be adhered
- Bridging nutritional gaps in developing nations



Conclusion and Future Work

- Transformer model trained on word level tokenization and nutrition specific texts generated much more coherent and relevant output than on character level
- Explore implementation of emerging technologies such as Multi- Head Latent Attention mechanism, Retrieval-augmented generation and Long Context Language Modeling
- Partnership with NGO, INGO and governmental bodies for open-source development and distribution

The background features abstract geometric elements: a large dark blue triangle in the bottom-left corner, a smaller dark blue triangle in the bottom-right corner, and a thin light blue diagonal line extending from the top-left towards the bottom-right.

**Thank you for
your attention**